#CrecimientoProfesional
#AprendeConLosPioneros

Online | DMC Perú

DMC Perú | CAPACITACIÓN PROFESIONAL

Especialización en Big Data

SESIÓN I

Docente: Mg. Ing. Layla Scheli

CAPACITACIÓN PROFESIONAL

PERFIL PROFESIONAL

Mg. Ing. Layla Scheli

https://www.linkedin.com/in/laylascheli/

- ✓ Profesional de la carrera de Ingeniería de Sistemas
- ✓ Master en Big Data y Business Intelligence
- ✓ Especialista en Tecnologías de la Informacion
- ✓ + 7 años de experiencia en áreas de Business Intelligence, Big Data, Software Factory y consultoras de TI a nivel nacional y extranjero.

REGLAS

Se requiere **puntualidad** para un mejor desarrollo del curso.

Para una mayor concentración **mantener silenciado el micrófono** durante la sesión.

Las preguntas se realizarán **a través del cha**t y en caso de que lo requieran **podrán activar el micrófono**.

Realizar las actividades y/o tareas encomendadas en **los plazos determinados**.

Identificarse en la sala Zoom con el primer nombre y primer apellido.

CAPACITACIÓN PROFESIONAL

MALLA CURRICULAR

INTRODUCCIÓN A BIG DATA

- Conceptos de Big Data.
- La filosofía de Big Data: Las 5V.
- Big Data como marco de trabajo.
- Capas conceptuales.
- Arquitectura conceptual.
- Componentes tecnológicos disponibles.
- Arquitectura tecnológica.

ALMACENAMIENTO DISTRIBUIDO CON APACHE HADOOP 1

- Tecnologías batch sobre Big Data
- Entendiendo ETL, ELT.
- Trabajando de manera distribuida sobre un clúster
- Introducción a Hadoop (Onpremise y Cloud).
- Hadoop como ecosistema de almacenamiento.

ALMACENAMIENTO DISTRIBUIDO CON APACHE HADOOP 2

- HDFS como motor de almacenamiento
- YARN como gestor de recursos
- Transformación de datos con Apache Hive.
- Tablas externas, Particiones dinámicas, estáticas.
- SQL sobre MapReduce
- Entorno web con Apache Hue.

PROCESAMIENTO DISTRIBUIDO CON APACHE SPARK

- Introducción a Spark.
- APIs y Funciones con PySpark.
- Extracción de datos.
- Transformación de datos.
- Dataframe, RDDs.

PATRONES DE ACCESO EN DATA LAKE

- Introducción a Datalake.
- Definición de Capas del datalake.
- Poblamiento de capa Row.
- Poblamiento capa Staging.
- Poblamiento capa Analytics.
- Entorno de desarrollo en Databricks.
- Entorno de desarrollo en Jupyter con Dataproc – Google Cloud Platform.

PROCESAMIENTO DE DATOS EN REAL TIME

- Introducción a Real time.
- ¿Streaming, real time, near real time o micro batch?
- Arquitectura general para proyectos real time.
- Procesamiento real time con Spark Streaming.
- Procesamiento real time en Cloud.

BIG DATA EN CLOUD

- Infraestrutura Cloud vs OnPremise.
- Instalación de un clúster de Big Data.
- Servicios de Big Data en Azure.
- Servicios de Big Data en AWS.
- Servicios de Big Data en GCP.
- Despliegue de infraestructura sobre cloud.

PROYECTO INTEGRADOR

- Desarrollo del proyecto en equipos.
- Presentación Final.

Online | DiHiC Perú

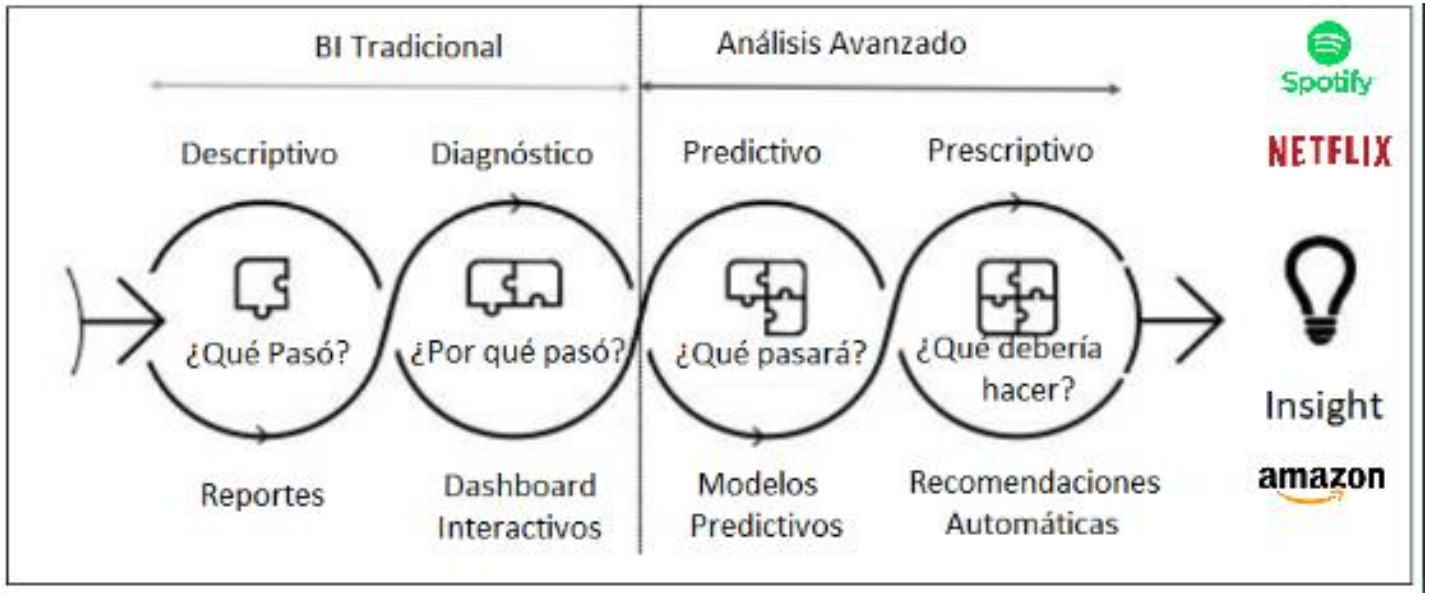
- ESPECIALIZACIÓN EN -

BIG DATA

#AprendeDesdeCasa
#AprendeConLosPioneros

Evolución de los datos

En el análisis avanzado de datos, a las compañías les interesa predecir el futuro y obtener recomendaciones automáticas.



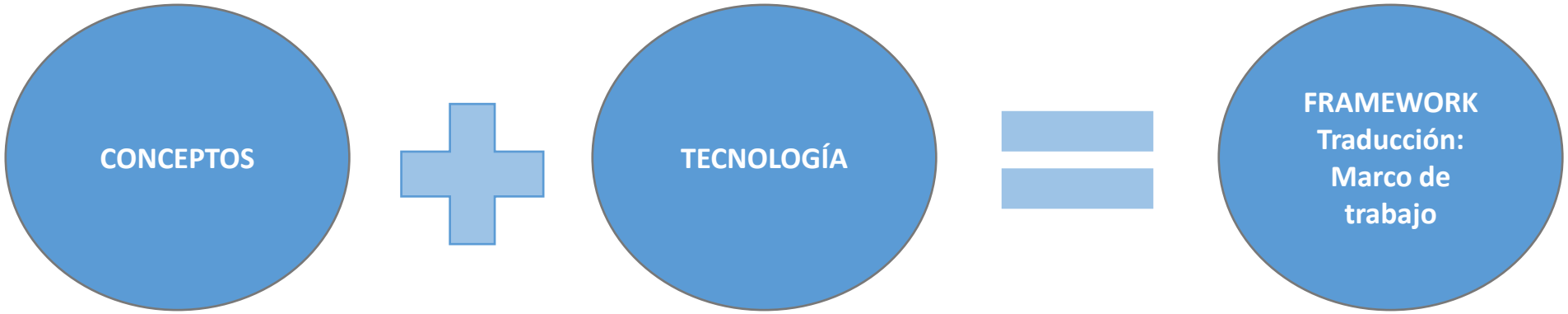
BIG DATA

Es un marco de trabajo **(conceptos + tecnologías)** que permite **procesar grandes volúmenes** de datos, de **diferentes estructuras o con carencia de estas**, que pueden **variar en el tiempo**, a **grandes velocidades** y que generen **valor** al negocio.

OBJETIVO FUNDAMENTAL DEL BIG DATA

1. **Reducir los tiempos** de procesamiento
2. **Integrar todas las fuentes** de datos disponibles
3. **Reducir los costos** de hardware
4. **Reducir el uso de recursos** computacionales
5. **Crecer fácilmente** en potencia computacional
6. **Aumentar la exactitud** en los cálculos
7. **Potenciar otras tecnologías** y marcos de trabajo

¿BIG DATA COMO FRAMEWORK?

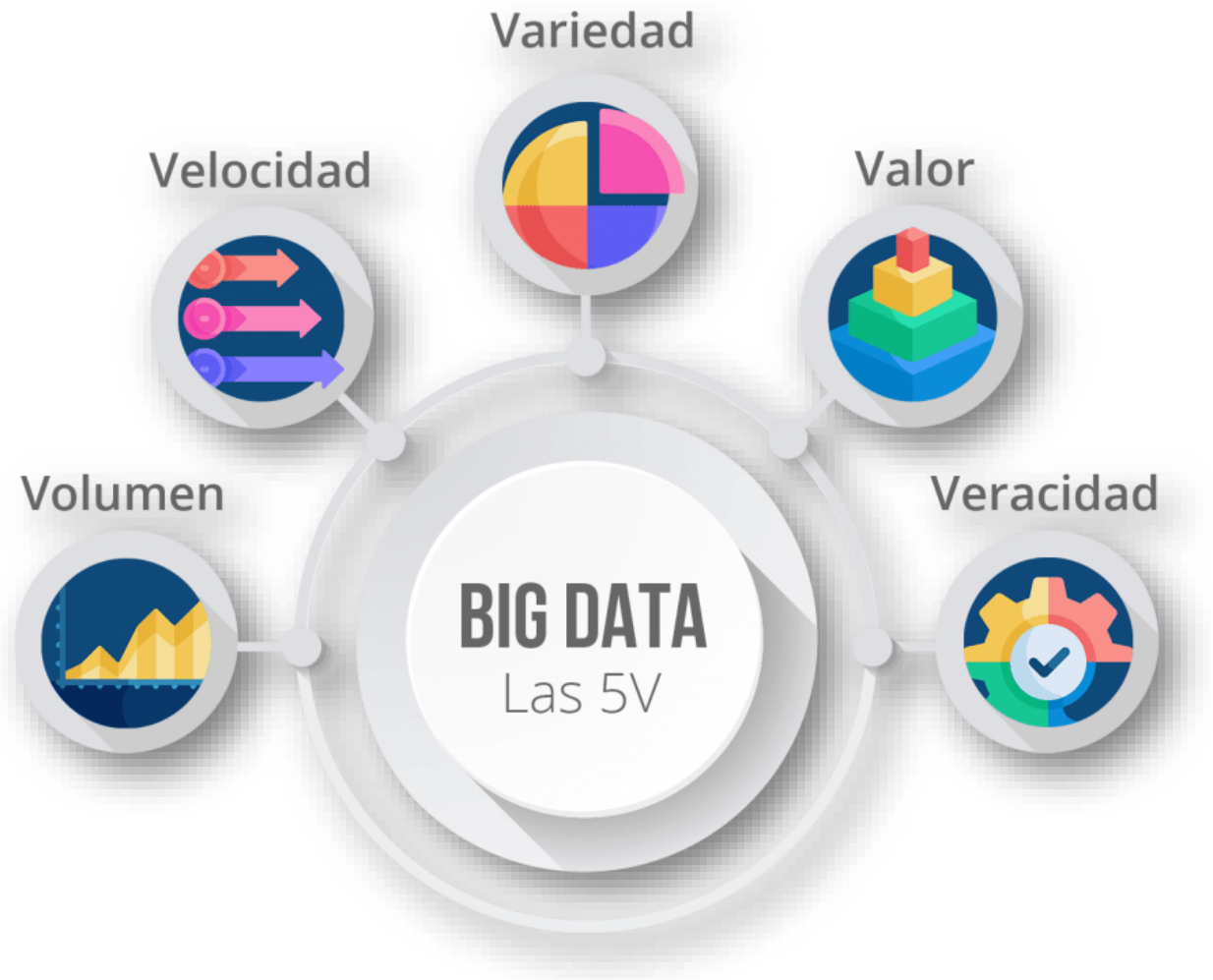


- Las 5vs
- Clúster computacional
- Paralelización
- Escalabilidad
- Alta Disponibilidad
- Seguridad
- Gobierno
- Patrones de diseño

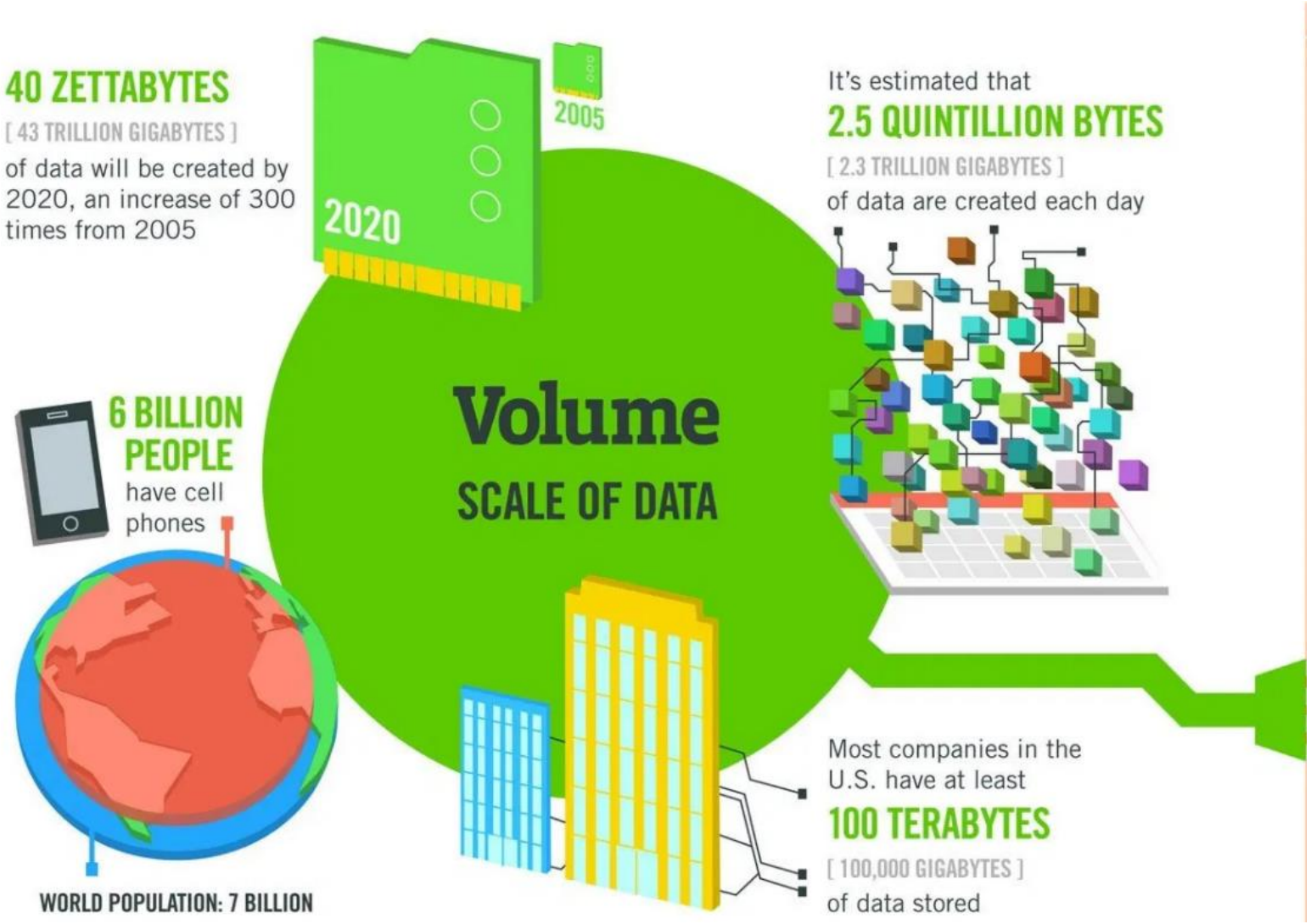
- Hadoop
- Hive
- Hbase
- Spark
- Kafka
- Cassandra
- Lenguajes de Programación
- AWS, Azure, GCP
- Y más.

BIG DATA

LAS V DEL BIG DATA



Volumen

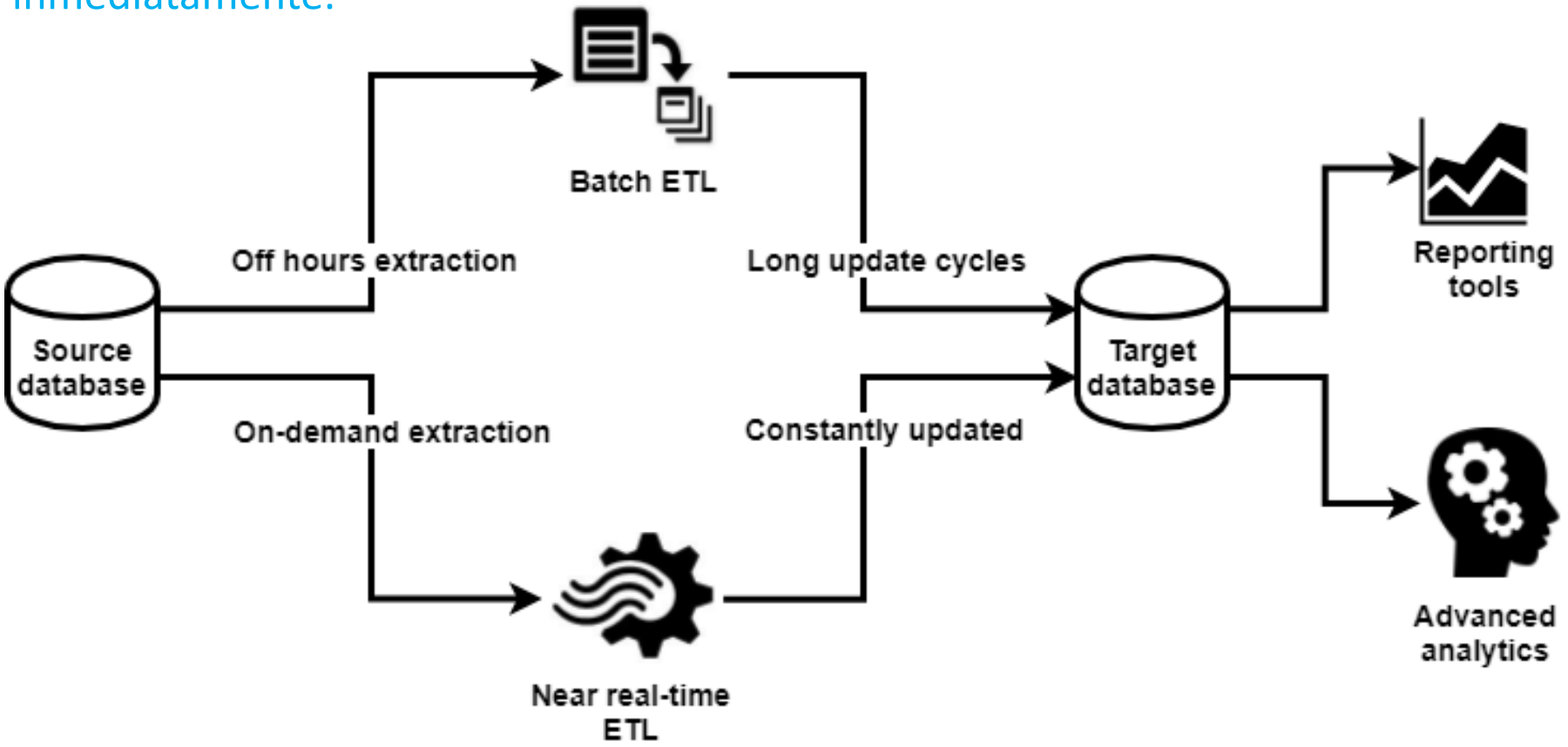

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2005
2020
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
Volume
SCALE OF DATA
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

UNIDADES DE MEDIDAS DE ALMACENAMIENTO

Medida	Simbologia	Equivalencia	Equivalente en Bytes
byte	b	8 bits	1 byte
kilobyte	Kb	1024 bytes	1 024 bytes
megabyte	MB	1024 KB	1 048 576 bytes
gigabyte	GB	1024 MB	1 073 741 824 bytes
terabyte	TB	1024 GB	1 099 511 627 776 bytes
Petabyte	PB	1024 TB	1 125 899 906 842 624 bytes
Exabyte	EB	1024 PB	1 152 921 504 606 846 976 bytes
Zetabyte	ZB	1024 EB	1 180 591 620 717 411 303 424 bytes
Yottabyte	YB	1024 ZB	1 208 925 819 614 629 174 706 176 bytes
Brontobyte	BB	1024 YB	1 237 940 039 285 380 274 899 124 224 bytes
Geopbyte	GB	1024 BB	1 267 650 600 228 229 401 496 703 205 376 bytes

VELOCIDAD

- **Procesamiento en batch:** Cuándo puede un tiempo prolongado, para el procesamiento.
- **Procesamiento en Near Real Time:** Cuando necesite que la información se procese inmediatamente.



Variedad

FUENTE DE DATOS

BD Transaccionales

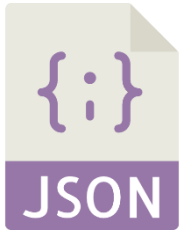
DATA ESTRUCTURADA

Misma estructura para todos los registros

Datos Real Time

DATA SEMI -ESTRUCTURADA

Cada registro tiene su propia estructura



Sensores

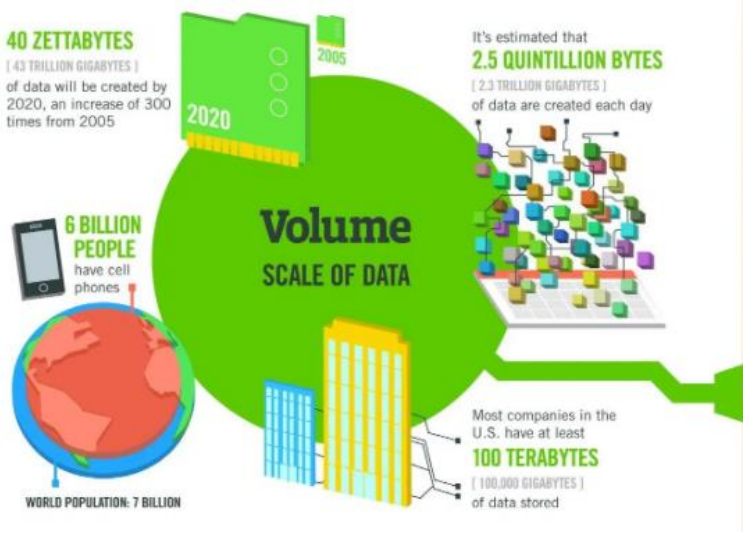
DATA NO ESTRUCTURADA

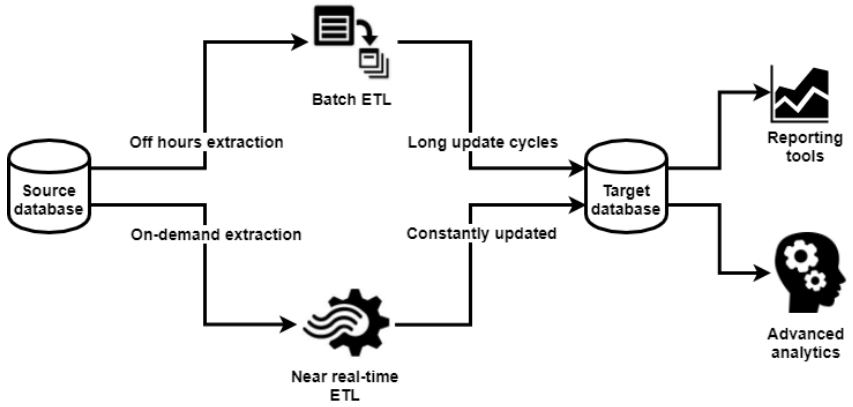
No tiene un estructura ni registro

Veracidad

- Se refiere al sesgo, el ruido y la alteración de datos.
- Los responsables de los proyectos de big data deben preguntarse si los datos son fiables y adecuados para los propósitos de análisis y necesidades de las organizaciones.

Valor






Aumentar ganancias

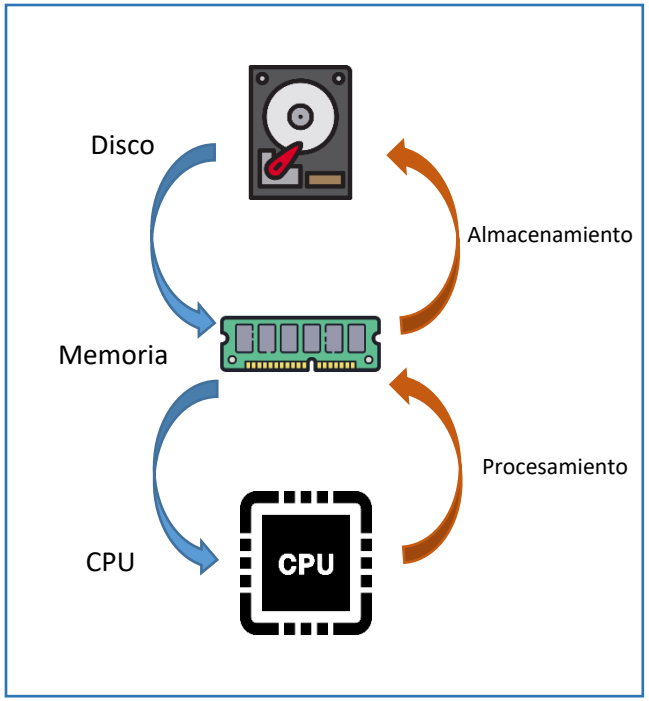

Diminuir pérdidas


Ahorro de tiempo


Incremento de satisfaccion

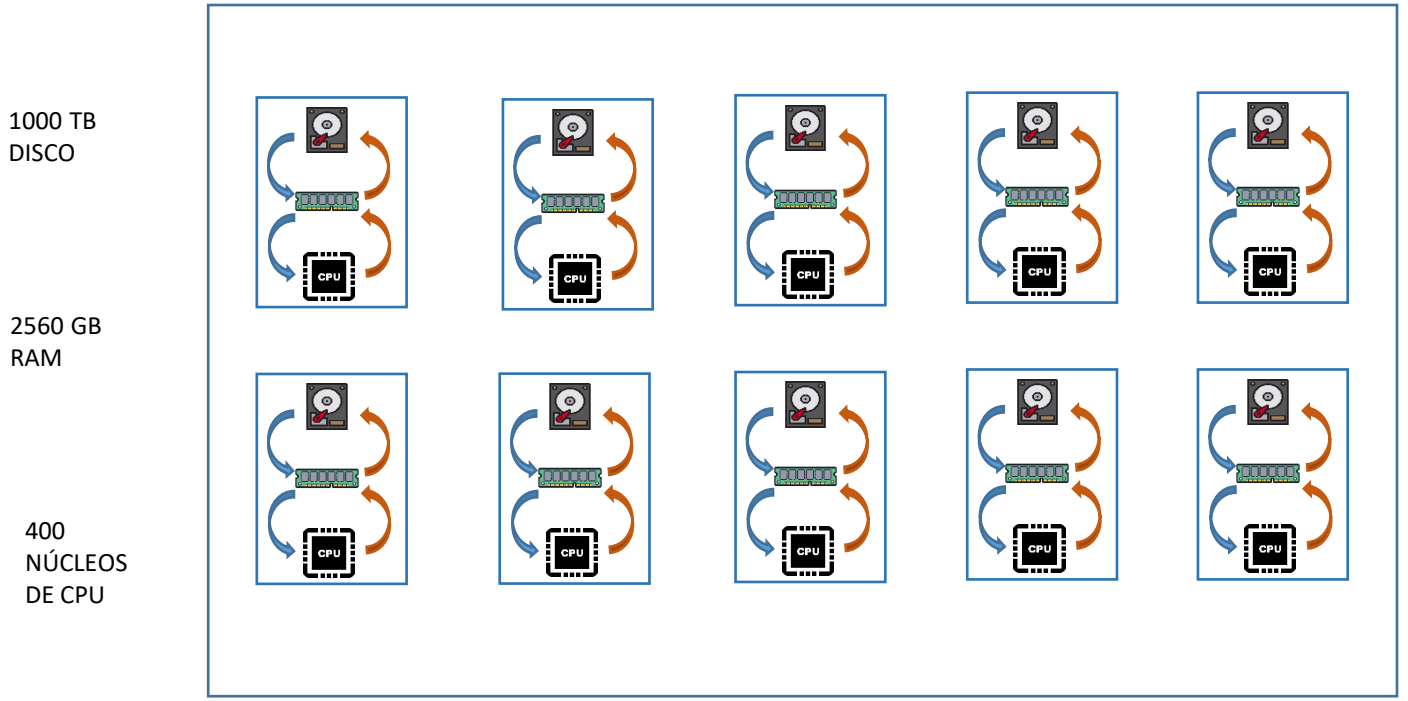
CLUSTER COMPUTACIONAL

¿Cómo trabaja una computadora?



SERVIDOR COMÚN

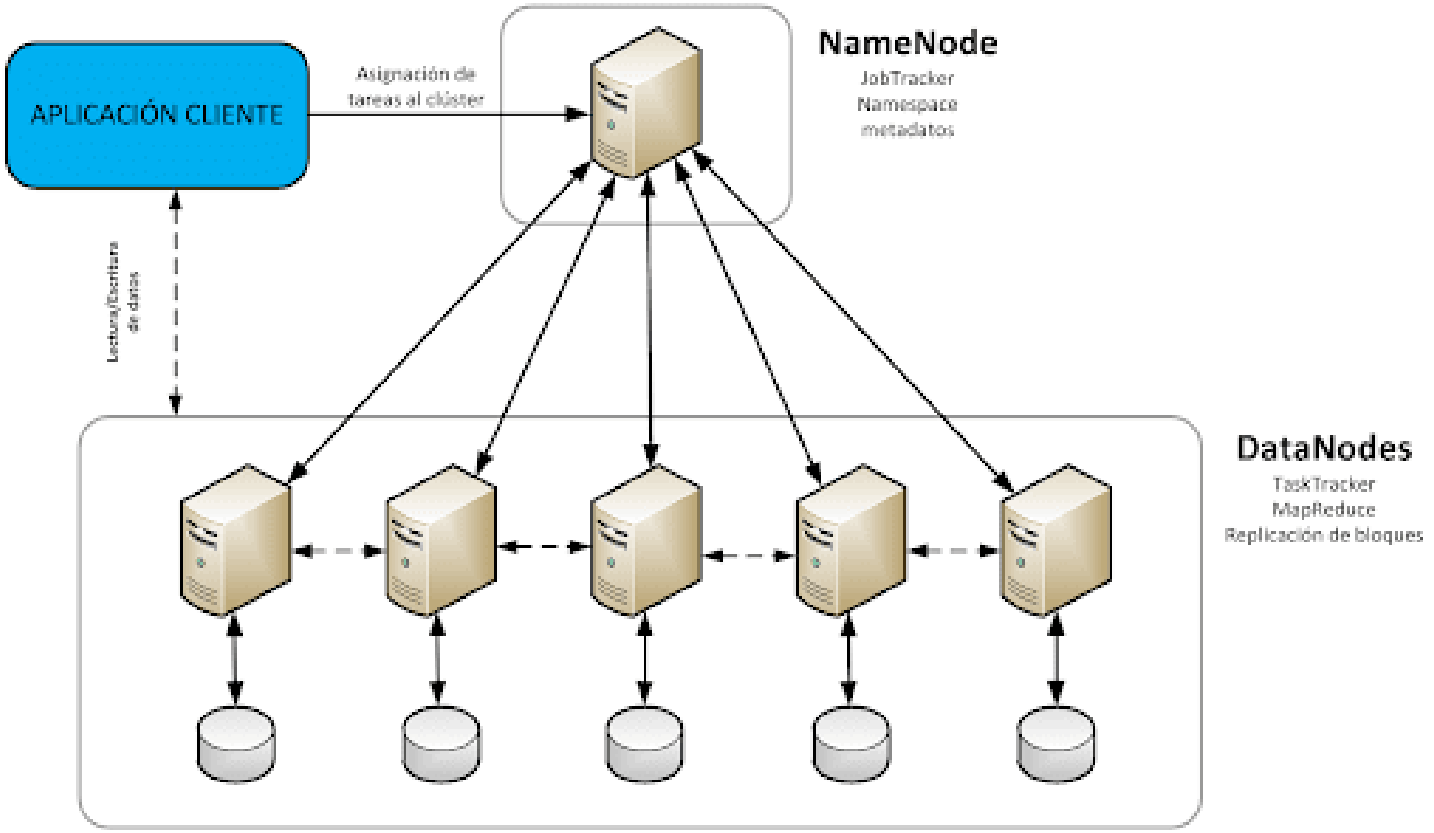
¿Cómo trabaja un clúster?



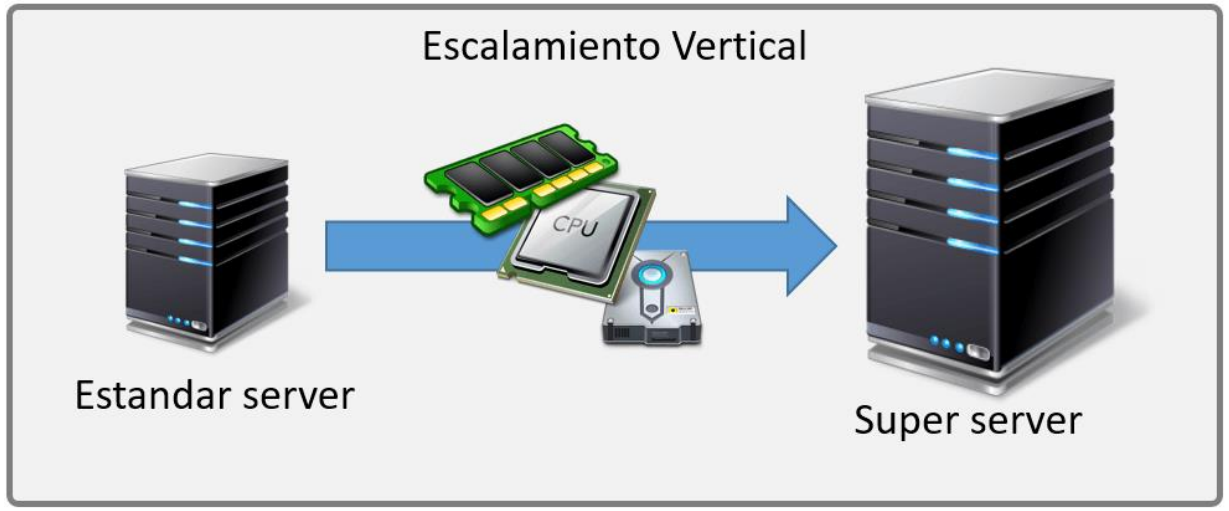
Un cluster es la suma de los recursos computacionales de los servidores que lo conforman, es como si tuviésemos una “super computadora”

PARALELIZACIÓN

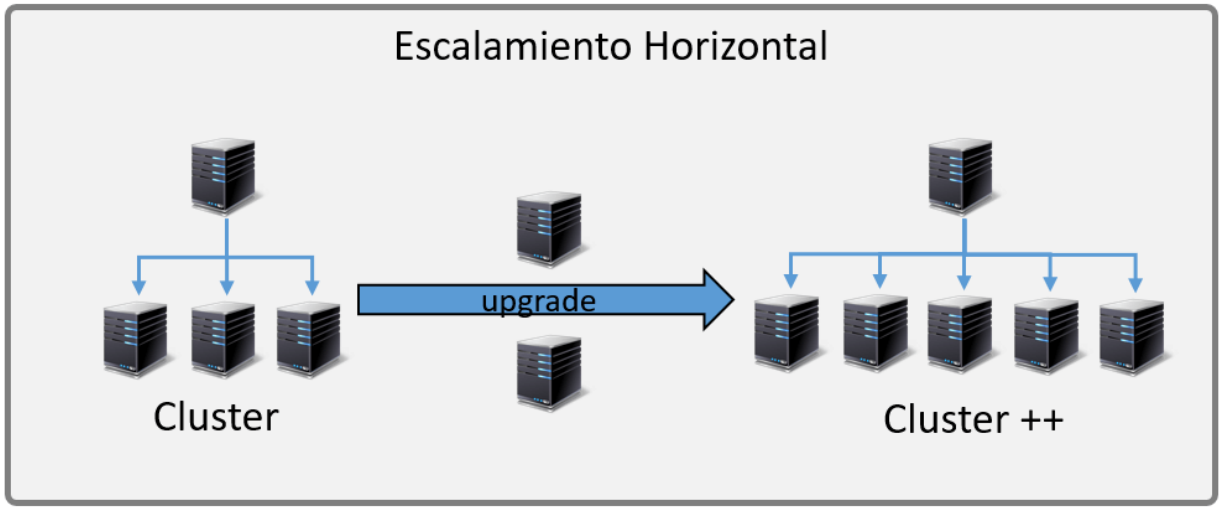
Paralelizar los datos en diferentes máquinas no es más que dividir los datos en archivos más pequeños. Estos archivos más pequeños son enviados cada uno a una máquina diferente. De esta forma, cada máquina procesará una pequeña parte del fichero inicial en lugar de analizar el fichero completo.



ESCALABILIDAD

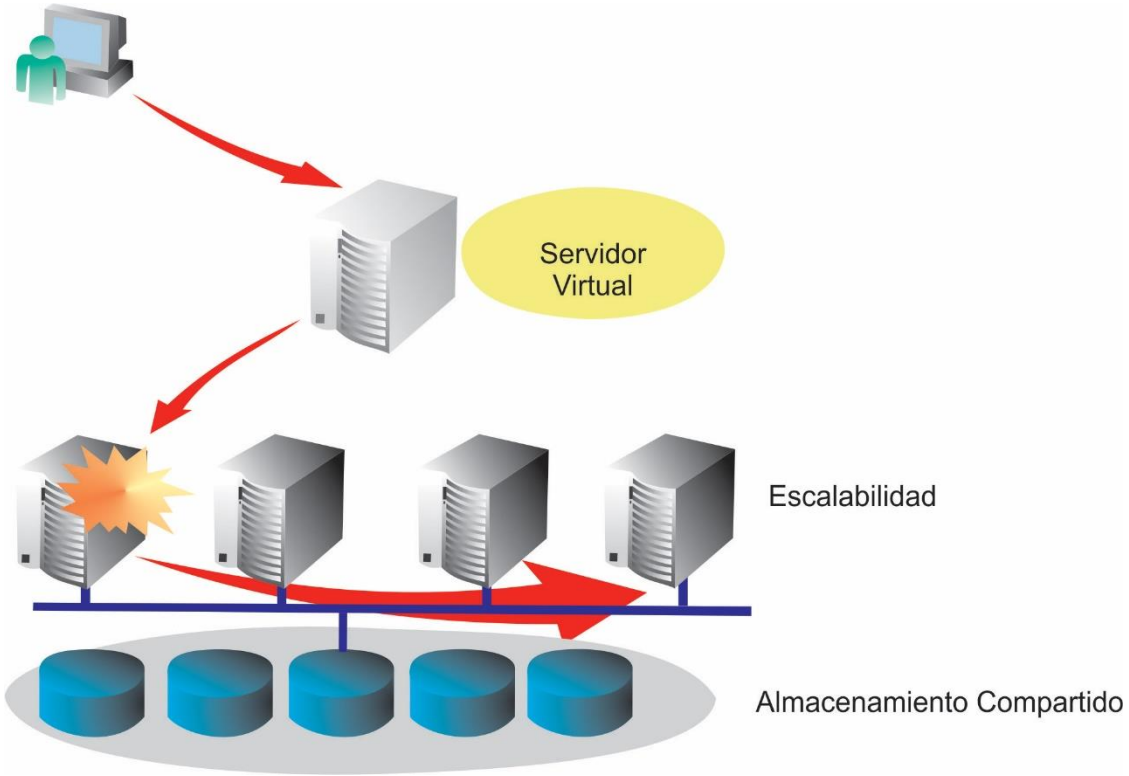


Enfoque Clásico



Enfoque Big Data

ALTA DISPONIBILIDAD



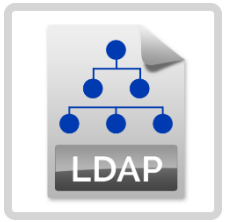
Índice de disponibilidad	Duración del tiempo de inactividad
97%	11 días
98%	7 días
99%	3 días y 15 horas
99,9%	8 horas y 48 minutos
99,99%	53 minutos
99,999%	5 minutos
99,9999%	32 segundos

https://cloudharmony.com/status-1year-of-compute-and-storage-group-by-regions-and-provider

SEGURIDAD

SEGURIDAD DE ACCESO REMOTO

Autenticación de Usuarios



CLUSTER

SEGURIDAD EN CLOUD

Tokenización

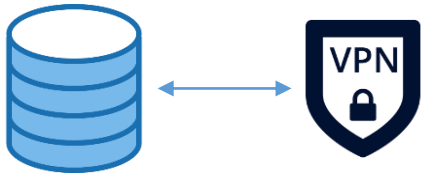
DNI	SALARIO
a94bf42190cd12001a4	1000
112833ac48b83dd2e9f	3500
b29481000c03e0cd00d	10500

Encriptación

SEGMENTO CLOUD

SEGURIDAD DE SERVICIOS

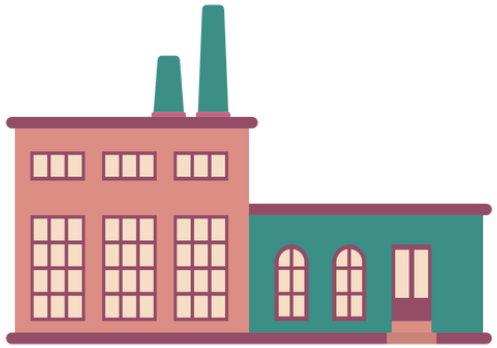
Fuentes de datos



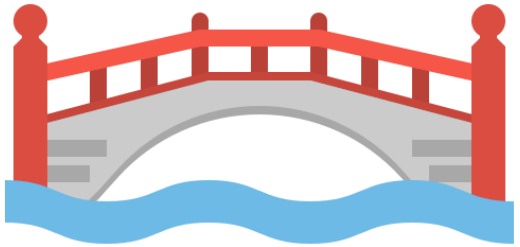
SEGURIDAD DE DATOS EN MOVIEMIENTO

PATRONES DE DISEÑO

Los **patrones de diseño** son soluciones habituales a problemas que ocurren con frecuencia en el diseño de software. Son como planos prefabricados que se pueden personalizar para resolver un problema de diseño recurrente en tu código.



PATRONES CREACIONALES



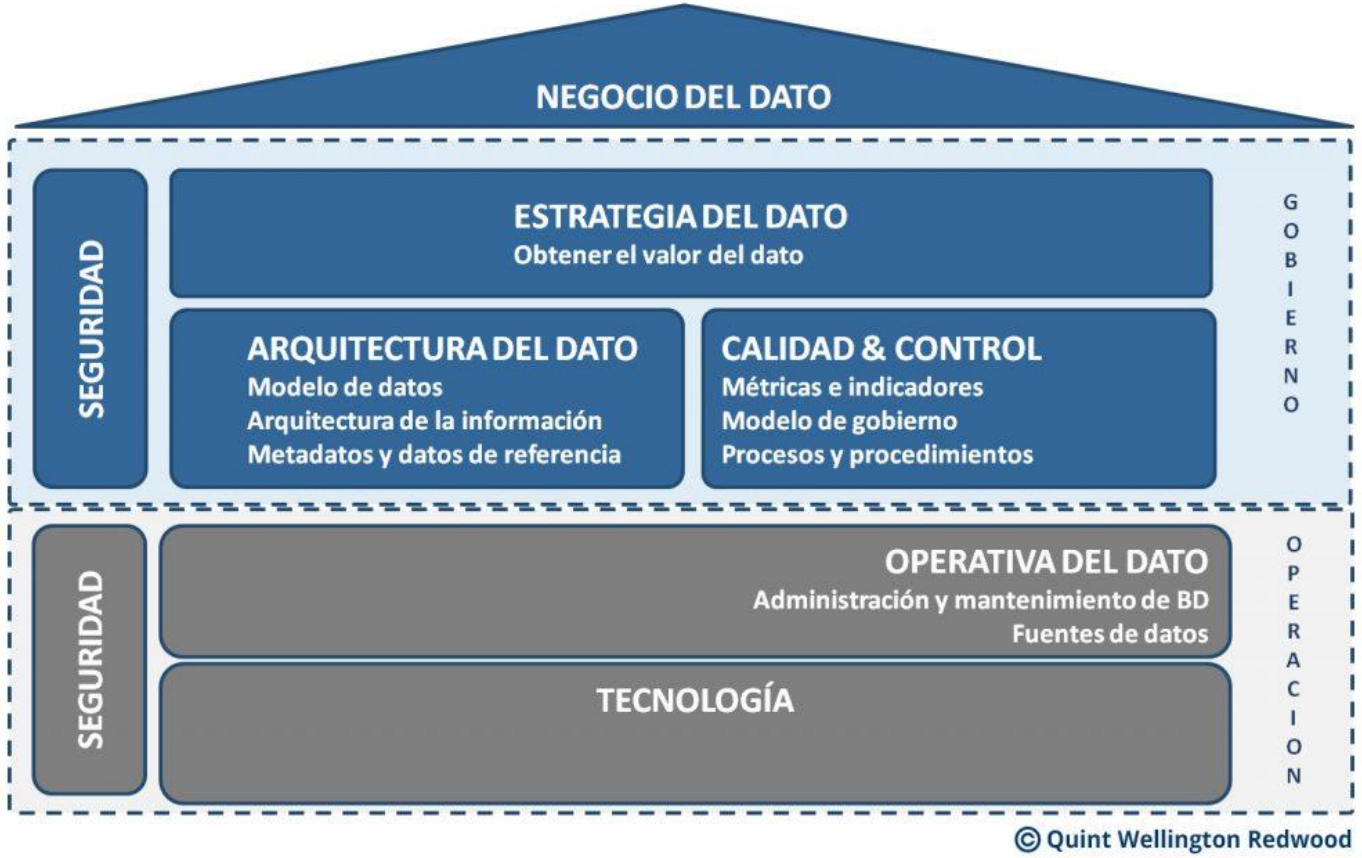
PATRONES ESTRUCTURALES

PATRONES DE COMPORTAMIENTO

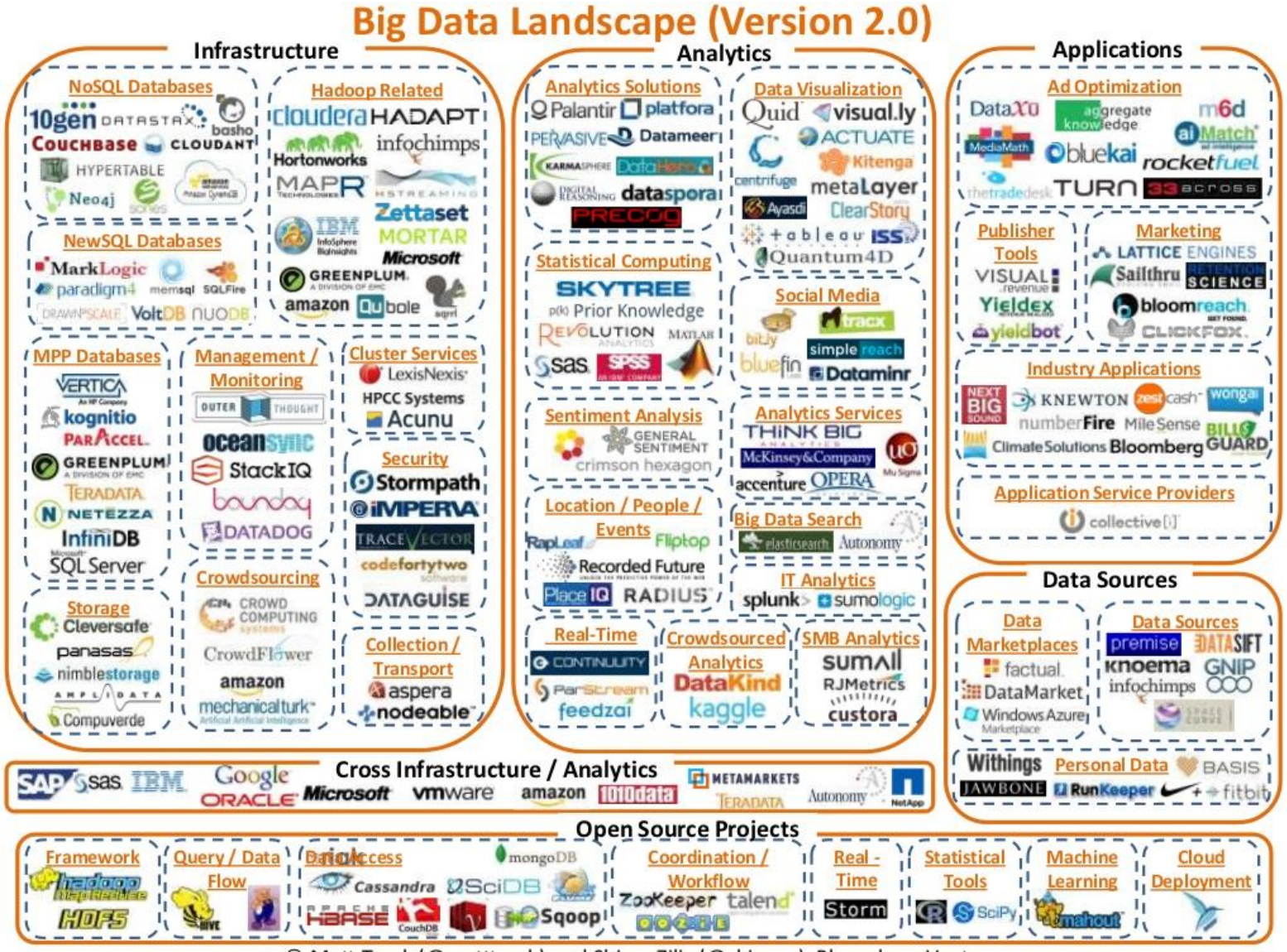
https://refactoring.guru/es/design-patterns/what-is-pattern

GOBIERNO

Es orquestación formal de gente, procesos y tecnología para permitir a una organización potenciar los datos como un activo de la compañía.

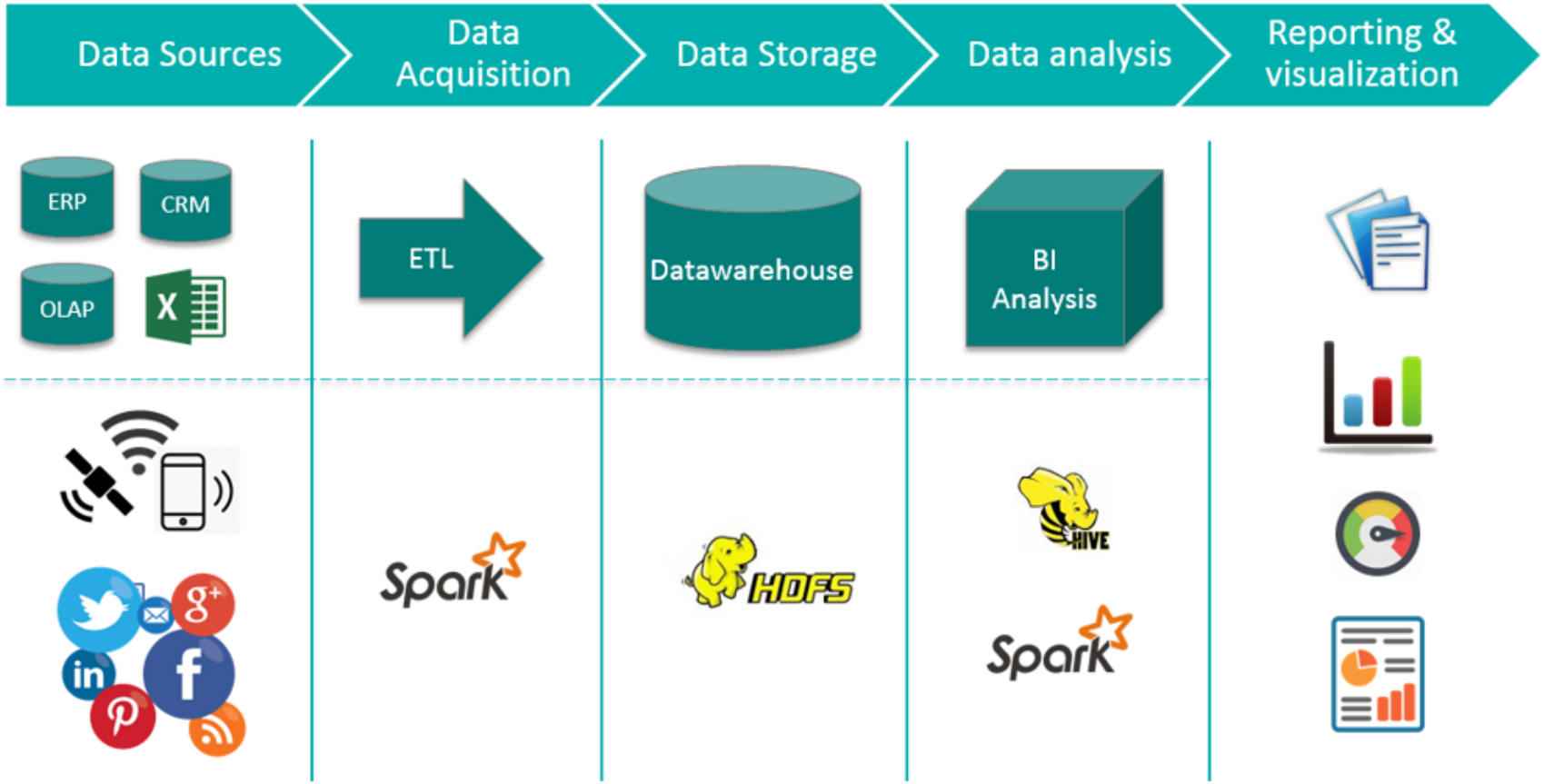


TECNOLOGÍAS



© Matt Turck (@mattturck) and ShivonZilis (@shivonz) Bloomberg Ventures

¿COMO **FUNCIONA** EL **BIG DATA**?



CAPACITACIÓN PROFESIONAL

LENGUAJES DE PROGRAMACIÓN

- Orientado a datos estructurados
- Aprendizaje fácil
- Estándar para consultas a bases de datos

- Orientado a la programación funcional
- Interpretado, no tipado y muy flexible
- **Librerías casi para todo**

- Orientado a la **programación de objetos**
- Compilado y tipado
- **Aprovecha muy bien los servidores con grandes recursos computacionales**

- Orientado a la programación funcional y de objetos
- Compilado y no tipado
- **Fork de Java**

- Orientado a la estadística
- Interpretado
- **Gráficos avanzados simples de realizar**

NUEVOS ROLES

- Data Engineer
- Data Scientist
- Data Expert
- Data Governace
- Data Architect
- Data Analyst
- Data Quality
- Chief Data Officer (CDO)

Volumen


aws
CERTIFIED
Big Data
Specialty


Google Cloud Certified
PROFESSIONAL
DATA ENGINEER


Microsoft
CERTIFIED
AZURE
DATA ENGINEER
ASSOCIATE


aws
certified
Machine
Learning
Specialty

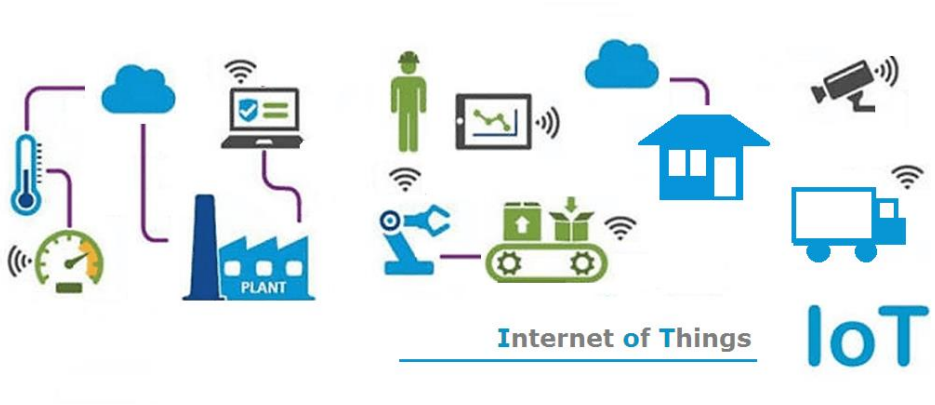

cloudera
Cloudera hereby recognizes that
Yung-Kuei Chen
has successfully completed all requirements
and is hereby certified as a
CCA Spark and Hadoop Developer
CCA
cloudera


PROFESSIONAL CERTIFICATE
CertiProf
Professional Knowledge
BIG DATA
BDPC


databricks
CERTIFIED
DEVELOPER
APACHE SPARK™ V2.X
2018


HORTONWORKS
HDP CERTIFIED
SPARK
DEVELOPER

TENDENCIAS



PREGUNTAS

1. ¿Qué es Big Data?
2. ¿Cuál es el objetivo del Big Data?
3. ¿Cuáles son las 5V?
4. ¿Qué es un clúster computacional?
5. ¿Qué es la paralelización?
6. ¿Qué es la escalabilidad?
7. ¿Qué tipos de tecnologías existen en el Big Data?
8. ¿Qué lenguaje de programación es mejor para programar sobre Big Data?
9. ¿Por qué se prefiere un proceso micro-batch sobre uno real time?

CAPACITACIÓN PROFESIONAL

LABORATORIO 01

CREACIÓN DE UN CLUSTER BIG DATA CON DATA PROC - GCP

DIIC Perú | CAPACITACIÓN **PROFESIONAL**

Online | DIIC Perú